DOCUMENT RESUME

ED 244 986 TM 840 308

AUTHOR Mayberry, Paul W.
TITLE A Study of Item Bias for Attitudinal Measurement Using Maximum Likelihood Factor Analysis.
PUB DATE Apr 84
NOTE 46p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 24-26, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Adults; Analysis of Covariance; *Attitude Measures; Data Analysis; *Factor Analysis; *Maximum Likelihood Statistics; Research Methodology; *Test Bias; Testing Problems; *Test Items
IDENTIFIERS *Inventory of Personal Investment

ABSTRACT

A technique for detecting item bias that is responsive to attitudinal measurement considerations is a maximum likelihood factor analysis procedure comparing multivariate factor structures across various subpopulations, often referred to as SIFASP. The SIFASP technique allows for factorial model comparisons in the testing of various hypotheses concerning the interitem covariance matrices. To demonstrate the utility of this technique, the Inventory of Personal Investment (IPI) was administered to three groups: academics, management, and college students. The basic hypothesis was that academics, management, and students with identical standing on any one of the seven latent incentives scales would have identical expected observed scores on the same scale, i.e., an identical factor structure would exist across the groups. The results showed that consistencies within the biased items can be identified, and that 13 of the IPI items did not provide equivalent measurement across the three subpopulations of this study. The maximum likelihood factor analysis procedure provided many advantages over the classical test theory and item response theory approaches for the detection of item bias in attitudinal instruments. (BW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED244986

A STUDY OF ITEM BIAS FOR ATTITUDINAL MEASUREMENT USING MAXIMUM LIKELIHOOD FACTOR ANALYSIS

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

P. W. Mayberry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paul W. Mayberry

Measurement and Research Division
Office of Instructional Resources
307 Engineering Hall
1308 W. Green
University of Illinois at Urbana-Champaign
Urbana, IL 61801

TM 810 308

Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA, April 1984. Comments are welcome.

A STUDY OF ITEM BIAS FOR ATTITUDINAL MEASUREMENT

USING MAXIMUM LIKELIHOOD FACTOR ANALYSIS

Research in the achievement domain provides strong support for the claim that item response theory (IRT) provides a theoretically rigorous procedure for detecting item bias, much more so than its classical test theory counterparts (Lord 1980, 1977, Wright 1977, Scheuneman 1979, Ironson & Subovlak 1979). While bias detection techniques developed for achievement measures have been implemented for the measurement of attitudes, the transfer between the two areas is not always compatible. Being such a powerful theory of measurement, IRT relies on a strong adherence to its underlying assumptions; in particular, unidimensionality. Unidimensionality of the latent space is not a realistic assumption with most attitudinal instruments which are usually designed to measure an entire range of constructs. Efforts to produce qualifying unidimensional scales often result in too few items per scale, which in turn distorts item parameter estimates. Also, current parameter estimation procedures restrict the scoring of item responses to being dichotomous. Loss of information is inevitable in the transformation from a multiple response format to a binary classification of the item's content. Finally, sample size requirements for stable parameter estimates often restrict the application of IRT.

A technique that is more responsive to these attitudinal measurement considerations is a maximum likelihood factor analysis procedure comparing multivariate factor structures acros various subpopulations, often referred to as SIFASP (Joreskog 1971). This approach readily accomodates polychotomous item responses while being able to converge for a smaller number of items and a smaller sample size. Also, the factorial model is able to function in a multidimensional latent space as well as perform simultaneous comparisons of

two or more groups.

The SIFASP technique allows for factorial model comparisons in the testing of various hypotheses concerning the interitem covariance matrices. These hypotheses postulate the factor structure for various subpopulations by making constraints on the parameter estimates and holding these estimates invariant across the subgroups. If the usual assumptions of the factor analysis model are made (i.e., unique variables are mutually uncorrelated and uncorrelated with the common factors), the variance-covariance matrix of the observed variables can be expressed as:

$$(1) \quad \hat{\Sigma} = \Lambda_y \Phi \Lambda_y' + \Psi^2 .$$

where Λ_y is the pattern matrix, Φ is the interfactor correlation matrix, and Ψ^2 is a diagonal matrix of the uniquenesses.

The basic premise of this "equivalence of measurement" analysis is based on Meredith's (1964) application of Lawley's (1943-44) selection theorem with respect to the covariance matrix described in equation 1: "when the various populations are derivable as subpopulations from a parent population under selection on some external variable, there is a factor pattern that is invariant over populations" (Joreskog 1971, p. 409). It is <u>not</u> required that the subpopulations be determined by selection on some external variable as long as the groups are clearly defined and independent, in addition to the selection variable being independent of the unique variables. As a result of Meredith's proof, the expected observed scores for individuals with identical factor scores (i.e., equal standing on the latent trait) can be translated as being equivalent regardless of group membership.

In any analysis of factor structure, the central issue is determining the

number of significant common factors represented in the observed correlation matrix. Joreskog (1967) developed a large sample χ^2 test for the maximum likelihood procedure. A significant χ^2 is interpreted as implying that one or more common factors is required to explain the covariances of the observed variables (or possibly there is model misfit -- i.e., the variables are not multivariate normal). McGaw and Joreskog (1971) warn against the strict adherence to the χ^2 statistic because of its sample size dependency. Rather, Joreskog and Sorbom (1981) suggest that the χ^2 value be used as a measure of the "goodness of fit" for the hypothesized factor model. Large χ^2 values are indicative of bad fitting models and small χ^2 values of good fitting models. The degrees of freedom for the factor model can be used to gauge the magnitude of the χ^2 values. This approach is especially applicable to comparisons of different factor models by using a ratio of change in χ^2 to change in degrees of freedom (Kroonenberg & Lewis 1982).

Joreskog and Sorbom (1981) provide two other measures of the goodness of fit for the factor model. The first of these is called the goodness of fit index (GFI). The GFI is independent of sample size and is relatively robust to departures from normality by the observed variables. An alternative to the GFI that uses mean squares instead of total sum of squares in its calculations is the adjusted GFI. The AGFI takes the degrees of freedom for the factor model into account when computing the statistic. Both measures are an indication of the relative amount of variance and covariance jointly accounted for by the factor model and should range in value from zero to one. "Unfortunately, the statistical distribution (of GFI and AGFI) is unknown, even under idealized assumptions, there is no standard to compare it with" (Joreskog & Sorbom 1981, p. I.41). The second measure of goodness of fit is the root mean square residual (RMR), an index of the average of the residual variances and covariances. The value of RMR can only be interpreted with respect to the

observed variances and covariances, and generally the smaller this value, the better the fit.

While each of these statistics provide pertinent information concerning the overall fit of the factor model, in no way do they reflect the quality of the factor model (rather, only an assessment of the overall fit of the model to the sample data). It is possible even if the indices show that the overall fit is good that the underlying relationships of the model may be poorly determined. To detect such a possibility it is important to examine the residual variances and covariances ($\underset{\sim}{S} - \underset{\sim}{\Sigma}$). Since $\underset{\sim}{S}$ is not always expressed as a correlation matrix, Joreskog and Sorbom (1981) calculate normalized residuals. This term is approximately a standard normal variable although it is correlated from cell to cell. The normalized residual serves as a flag of insufficient fit by the model if its value is larger than two for a particular element of the $\underset{\sim}{S}$ matrix. Specific corrections of either relaxing or constraining additional parameters of the model can be implemented according to "modification indices" (Joreskog & Sorbom 1981). The index is associated with the derivatives of the fitting function for all fixed and constrained parameters and is distributed as a χ^2 with one degree of freedom. If the fixed parameter with the largest modification index is freed, a decrement of <u>at least</u> the magnitude of the index will result in the associated χ^2. The process follows such that successive parameters with the largest index are freed one at a time until the change associated with χ^2 is no longer unacceptable.

METHODOLOGY

Instrument

Based upon the theory of personal investment, Braskamp and Maehr (1983) have developed an Inventory of Personal Investment ($_IP_I$). Analogous to financial investment, the theory states that individuals expend differing amounts of their time, talents, and resources in an attempt to maximize their overall satisfaction or feelings of importance. There are various personal incentives for investing oneself in some areas but not others. Seven of these incentives were studied here. A total of 838 adults completed the 61 item inventory. This sample was composed of three heterogenous groups: academics (N=322), management (N=248), and college students (N=268).

The $_IP_I$ was developed under the pretense of providing equivalent measurement across various groups and organizations although this has never been directly tested. The basic hypothesis was that academics, management, and students with identical standing on any one of the seven latent incentives scales would have identical expected observed scores on the same scale; i.e., an identical factor structure would exist across the groups.

Normality Assessment and Subscale Formation

The initial step in testing the equivalence of measurement for the $_IP_I$ was to determine the extent to which the inventory satisfied the requirement for normally distributed item responses. Therefore the correlation between each item response and its corresponding propensity variable, $\rho(X_i, L_i)$, was be determined for all sixty-one items (Drasgow & Dorans 1982). Dependent upon the obtained values, it might be necessary to form subscales of the items.

Item content was the most important criterion in the construction of the subscales. Given that the $_IP_I$ was composed of seven relatively distinct

scales, subscales were created within each of these "content" areas. Thurstone (1947) stated that at least three elements (items) should represent each factor of an instrument. This resulted in the formation of at least twenty-one subscales to represent the seven scales of the $_{I}P_{I}$. Within each $_{I}P_{I}$ scale, it was important to examine the distributional characteristics of the item responses in order to maximize the number of used response categories for the resulting set of subscales.

Subscales were used for all subsequent analyses regardless of whether or not the item responses were normally distributed. Such a strict adherence to the use of subscales as the unit of measurement was required even though item analysis was preferable. The use of subscales was necessary due to the fact that the $_{I}P_{I}$ was composed of so many items and the computional methods for determining maximum likelihood solutions were extremely time consuming when working with so many variables. If the subscale analyses showed that equivalence of measurement did not extend across all subpopulations, then the analysis would be repeated for the biased subscale but on an item level (if the items themselves have been determined to be normally distributed).

Determination of the Optimal Factor Solution

Once the twenty-one subscales were constructed for the seven $_{I}P_{I}$ scales, it was necessary to conduct a preliminary factor analysis. Such an analysis of the subscales determined the approximate factor structure as well as provided initial starting values for the maximum likelihood factor analysis. By inputting such values, the number of iterations required before convergence to the maximum of the likelihood function was greatly reduced. The correlation matrix for the subscales was computed and the diagonal elements replaced by squared multiple correlations of each subscale with the other twenty subscales. This correlation matrix was analyzed by a principal axis factoring procedure

for one to ten factors. Each solution was obliquely rotated to achieve simple structure. The eigenvalues resulting from each of these analysis were compared to pseudo random eigenvalues that were computed from a parallel analysis (Humphreys & Montanelli 1975) to determine the approximate number of common factors represented in the observed data. The pseudo random eigenvalues were calculated from a randomly generated correlation matrix with the squared multiple correlations of the subscales substituted in the diagonal. The point at which the plot of the real eigenvalues crossed the plot of the pseudo random eigenvalues determined the number of common factors for the subscales.

Based on the factor loadings that were calculated in the preliminary analysis, a maximum likelihood factor analysis was conducted to confirm these initial results. Joreskog and Sorbom's (1981) LISREL computer program was used to estimate all parameters for the maximum likelihood solutions. In order to obtain a unique factor solution from LISREL, it was necessary to constrain $\underline{m}^2$ parameters of the factor model (Joreskog & Sorbom 1979). This is achieved by fixing the diagonal elements of Φ at unity and $\underline{m}-1$ elements in each column of Λ_y at zero. To determine which elements to fix at zero, the largest loading in each column of Λ_y was found. All elements of the $\underline{m}-1$ corresponding columns were fixed at zero. Since the χ^2 statistic is not always appropriate for the assessment of the maximum likelihood solution, the Tucker-Lewis reliability coefficient was also calculated (Tucker & Lewis 1973). Based on these statistics along with the other goodness of fit indices provided by LISREL, the best maximum likelihood solution was determined.

The above factor solution was further refined in order to minimize overfitting of the factor model by fixing additional parameters of Λ_y at zero (Kroonenberg & Lewis 1982). One of the benefits of maximum likelihood estimation is that standard errors could be calculated for each parameter of the factor model. From the standard errors, a $\underline{t}$-statistic can be computed for

each element of Λ_y to determine whether it was statistically different from zero. Those t-statistics small in magnitude were likely candidates to be fixed at zero. Once these elements had been identified, the LISREL analysis was repeated to determine if the additional constraints seriously affected the fit of the factor model.

The χ^2 value for the simplified model (more parameter fixed at zero) should not substantially exceed the χ^2 for the original model given the number of imposed restrictions. If the ratio of change in χ^2 to the change in degrees of freedom was large, the additional constraints imposed on the factor model caused a departure by the factor model from the observed data. Therefore the number of constraints (number of parameters fixed at zero) must be relaxed. Using the modification indices, the exact misfitting parameter can be identified and accordingly be freed to provide a better fitting model. In summary, the factor model for which the parameters with t-statistics less than 2.58 (alpha = .05) in absolute value were fixed at zero served as the upper limit for the "optimal" factor model. Further refinements were made on the model by freeing the parameter with the largest modification index (only one parameter was freed at a time). This process was continued until subsequent freeing of parameters did not significantly decrease the χ^2. The resulting factor model would serve as the foundation for the analysis of educators versus middle management versus students.

SIFASP for the Subpopulations

The SIFASP analysis was a process of sequential hypothesis testing. Dependent upon the conclusions of the significance tests, the analysis either branched to another test or terminated. This progression of significance tests is given in Figure 1. The figure provides an overview for the SIFASP analysis and will serve as a useful reference for the following discussion. The null

hypothesis for each test was given along with the test statistic that was used to determine the conclusion for that analysis. The course of action was given for both significant and nonsignificant results.

Before the SIFASP was conducted, it was necessary to rescale the twenty-one subscales since their units of measurement vary across subscales due to differing numbers of items for each. Typically the variance-covariance matrices for the subpopulations are rescaled to correlation matrices, yet such a transformation would eliminate important group differences (McGaw & Joreskog 1971). An acceptable rescaling technique was based on a weighted pooled variance-covariance matrix, $\underset{\sim}{S}$:

$$(2) \quad \underset{\sim}{S} = \frac{1}{N - g} \sum_{i=1}^{g} (N_i - 1)\, \underset{\sim}{S}_i \, .$$

where $N = \sum N_i$ and $i = (1, 2, \ldots, k)$, the number of subpopulations.

A diagonal matrix of the reciprocals of subscale standard deviations was computed from $\underset{\sim}{S}$:

$$(3) \quad \underset{\sim}{D} = (\text{Diag}\, \underset{\sim}{S})^{-1/2} \, .$$

The original variance-covariance matrices was then rescaled by this diagonal matrix:

$$(4) \quad \underset{\sim}{S}_i^{*} = \underset{\sim}{D}\, \underset{\sim}{S}_i\, \underset{\sim}{D} \, .$$

The preliminary test in a SIFASP analysis was to determine if the rescaled variance-covariance matrices for the subpopulations were equivalent. If the matrices were not significantly different, then equivalence of measurement held

across the groups and the analysis stopped. Box (1949) developed a F test to test whether there was equality across these matrices. If the null hypothesis of equality was rejected, subsequent tests were performed to test the invariance property of Λ_y.

Dependent upon the significance of the F-statistic for the above test, a separate maximum likelihood factor solution would be computed for each subpopulation. These independent solutions followed the simplified optimal factor solution that was calculated for the entire sample; i.e., the same number of common factors was used and the same elements of Λ_y were held constant, fixed at zero, or set free to be estimated. The strategy was to compare this "unconstrained" factor model with a model in which Λ_y and Ψ^2 were required to identical across subpopulations.

Since the subpopulations were formed such that they were mutually exclusive and independent, the respective χ^2 values associated with the separate maximum likelihood solutions were also independent. The sum of the independent χ^2 values was also distributed as a χ^2 and would accordingly represent the χ^2 that resulted from letting Λ_y, Φ, and Ψ^2 vary across subpopulations. The χ^2 value for the "constrained" model was determined by a simultaneous factor analysis of each rescaled variance-covariance matrices, S_i^*. In this analysis, Λ_y and Ψ^2 were required to be the same for all subpopulations while the entire Φ matrix was set free to vary. By totally freeing Φ, an additional m constraints must be imposed on the Λ_y matrix to ensure that the maximum likelihood solution was unique (Joreskog & Sorbom 1979). Thus an additional non-zero element (the largest loading) was fixed in each column of Λ_y to account for the m constraints lost by freeing the diagonal of Φ. The difference in the χ^2 values for the two models was tested to determine if the requirement of invariance for Λ_y and Ψ^2 significantly affected the fit of the unconstrained factor model. If the resulting ratio

of χ^2 values for the test of educators versus middle management versus students was significant, the conclusion would be that constraining both Λ_y and Ψ^2 across the subpopulations caused an inordinate degree of variance in the factor model and thereby causing it to not sufficiently fit the observed data.

A follow-up analysis in which Ψ^2 was not constrained to be equal across all groups would test the invariance property of Λ_y. If the resulting ratio of χ^2 to degrees of freedom was small then the conclusion was that only the uniquenesses, Ψ^2, were not invariant across the groups and equivalence of measurement held. However, if the ratio was small, Λ_y was not invariant across the subpopulations and therefore there was no equivalence of measurement. The task then became one of identifying the exact subscale or subscales which were the locus of effect and repeating the analysis but on an item level.

Testing the Equality of Subscale Factor Loadings Across Subpopulations

The primary concern before conducting the test of equality of subscale factor loadings was that the estimated loadings be expressed in comparable metrics (Mulaik 1972). Such caution was essential to ensure that consistent differences among the estimated loadings were not deemed as being significantly different merely due to the fact that the loadings were based on dissimilar scales of measurement. To achieve comparable metrics for the estimated loadings, the variance-covariance matrix for each subpopulation was rescaled as was done earlier according to equations 2-4. These rescaled variance-covariance matrices then served as input for the factor analyses as opposed to using the original correlation matrices. In addition to rescaling the variance-covariance matrices, it was necessary to also express the factor variances in similar units. Drasgow and Kanfer (1983) suggested that the

variances (the diagonals of the Φ matrices) be fixed at the values that were obtained in the earlier simultaneous factor analysis for which the Θ^2 and Φ matrices were entirely free and Λ_y was held invariant across the groups. Having completed both of these steps, the estimated loadings resulting from the separate maximum likelihood factor analyses were essentially in analogous metrics and were therefore directly comparable.

The test statistic for the comparison of the estimated loadings across subpopulations was Wilks' (1932) lambda criterion. Applicable to any number of groups, the statistic was a ratio of determinants for the within groups and the total SSCP matrices. The intermediate step for determining lambda first involved calculating the within group SSCP matrix. Tatsuoka (1971) stated that W was defined as the sum of the separate SSCP matrices for each group. For the case of three groups, W was calculated as follows:

$$(5) \quad W = S_1 + S_2 + S_3 .$$

The SSCP matrix for each group was not directly provided by the LISREL output but the program did supply the essential components necessary to compute each matrix. It was important to note that the rank of each S_i matrix will be of order p, the number of loadings estimated for a particular subscale. Recall that many loadings were fixed at zero in order to reduce overfitting by the factor model and therefore these zero loadings were consistent over groups and would not be included in the analyses.

Factor loadings computed on the basis of maximum likelihood estimation procedures have the distinct property of being multivariate normally distributed. Therefore in large samples, it was possible to determine the sampling variance and covariances for each loading that was estimated. The sampling variances were calculated by squaring each standard error of the

estimated loadings. Likewise, the sampling covariances were the products of all pairwise standard errors multiplied by the accompanying correlation between the two loading estimates. Sampling variances and covariances were determined for each subpopulation:

$$(6) \quad \underset{\sim}{S}_i = \begin{pmatrix} SE_{i1}^2 & & & \\ SE_{i1}\,SE_{i2}\,r_{12} & SE_{i2}^2 & & \\ \vdots & \vdots & \ddots & \\ SE_{i1}\,SE_{ip}\,r_{1p} & SE_{i2}\,SE_{ip}\,r_{2p} & \cdots & SE_{ip}^2 \end{pmatrix}$$

The within group SSCP matrix was then determined by summing the individual SSCP matrices for each subpopulation.

The total SSCP matrix was also not directly provided by the LISREL output but rather could be computed as the sum of the within and between SSCP matrices. The between groups SSCP matrix was defined by:

$$(7) \quad \underset{\sim}{B} = (\bar{\underset{\sim}{X}} - \bar{\bar{\underset{\sim}{X}}})'(\bar{\underset{\sim}{X}} - \bar{\bar{\underset{\sim}{X}}}).$$

where $\bar{\underset{\sim}{X}}$ was a row vector containing the average of the $\underline{p}$ factor loadings across groups and $\bar{\bar{\underset{\sim}{X}}}$ was a scalar representing the grand mean of all factor loadings for that subscale.

Having determined the between groups matrix, this matrix was added to the within groups SSCP matrix to yield the total SSCP matrix. The ratio of the determinants of the within to the total resulted in a sample value of Wilks' lambda. The lambda criterion is an inverse function of an $\underline{F}$-ratio. For the case of three (K = 3) subpopulations, the function reduces to an exact $\underline{F}$-ratio (Tatsuoka 1971), expressed by R:

$$(8)\quad \text{For } K = 3 \qquad R = \frac{1 - \Lambda^{\frac{1}{2}}}{\Lambda^{\frac{1}{2}}} \cdot \frac{N - p - 2}{p} \sim F^{2p}_{2(N-p-1)} .$$

It is important to note that N in this calculation did not represent the sample size of the subpopulations but rather the number of estimated factor loadings (N = K * P) on which R was based. Such a test would be conducted for each subscale of the ${}_{I}P_{I}$. The strategy was to identify those subscales across which the factor loadings were not consistent. Having determined the subscales for which the loadings significantly varied across the subpopulations, the analyses turned to investigating the "biased" subscales but on the item level.

Analysis of Subscales Identified as Being Biased

The subscales that were identified as having inconsistent factor loadings across the subpopulations were then expanded to their original items. In this manner, the exact item or items which were the cause of the overall bias could be determined. It is important to recall that the item responses need to be approximately normally distributed to justify such an analysis. If the item responses were not sufficiently normally distributed, conclusions could only be based on the subscales analyses. Given that the responses were approximately normal, the procedure for the analysis of the items was essentially the same as that for the analysis of the subscales; only a few new restrictions applied.

For the ${}_{I}P_{I}$, the subscales would be formed of two to four items. As a result of expanding the biased subscales to the item level, the number of possible variables entering into the analysis might become too large for the LISREL program (recall that subscales were originally used in order to reduce the number of variables to a manageable level). So that the item analysis would not be overwhelming, the only variables entering into the item analyses were the original items from the biased subscales and a reference subscale for

each biased subscale.

As stated earlier, the analysis of the items and the appropriate subscales basically followed the procedures discussed for the SIFASP. First, a continuity assessment based on $\rho(X_i, L_i)$ for each item of the biased subscale was conducted to determine if an item analysis was justified. Assuming that the items were sufficiently normal, the subscales for inclusion in the analysis were then identified. Separate maximum likelihood factor solutions based on the correlation matrices of items-subscales for each subpopulation were computed. The starting values for the Λ_y, Φ, and Θ^2 matrices were the LISREL estimates determined in the previous maximum likelihood factor solutions for each subpopulation. The starting values of the Φ matrices were taken directly from the LISREL estimates determined for each subpopulation. For Λ_y and Θ^2, modifications were made of the LISREL estimate matrices to only include the appropriate subscales. The starting values for the items forming the biased subscale were the LISREL estimate for that subscale repeated for each item. Using the solutions from previous analysis as the starting values was quite acceptable, an exploratory factor analysis was not necessary to determine these values. Starting values were just that--starting values, the better these values, the fewer the number of iterations required before the maximum of the likelihood function was reached. The resulting χ^2 value for each of the solutions was then summed and represented the overall χ^2 for the unconstrained factor model in which Λ_y, Θ^2, and Φ were free to vary across subpopulations.

It was necessary to compute rescaled variance-covariance matrices for each subpopulation to use as input for the overall SIFASP. These matrices were rescaled according to equations 2-4 and included only the items and the appropriate subscales. In addition to rescaling these matrices, the diagonal elements of the Φ matrix were set free to be estimated. This required that an

additional m elements of the Λ_y matrix be fixed. The starting values for the SIFASP were taken from the previous SIFASP analysis in which Λ_y was held invariant. Again, modifications of the LISREL estimates were made so that only the appropriate subscales were included and that the starting values for the items reflected the estimates for the biased subscale. The χ^2 resulting from the SIFASP represented the degree of fit for the constrained factor model in which Λ_y was held invariant across groups. This χ^2 was then compared with the χ^2 for the unconstrained factor model. If the increase was significant, a test would be conducted to determine the invariance property of Λ_y for each item. It was almost certain that this test would be significant since it was based soley on the items of previously determined biased subscales.

Having confirmed the previous analyses that identified potentially biased subscales, the task was to identify those items with inconsistent across groups. Recall that it was essential that the factor loadings were expressed in comparable metrics. A common metric for the estimated loadings was achieved by first using rescaled variance-covariance matrices and then by fixing the factor variances at the values obtained in the earlier SIFASP for which Φ^2 and Φ were free and Λ_y was held invariant. Wilks' lambda criterion was then calculated for each item (not each subscale) that formed the biased subscale. For those items in which Wilks' lambda was significant, the conclusion was that the estimated factor loadings were not consistent over the groups.

Having identified the biased items from all of the items forming the biased subscale, it was necessary to locate the specific subpopulations over which Λ_y was not invariant. For the case of three groups, three pairwise comparisons were made for each biased item. The test statistic for such comparisons followed a χ^2 distribution with its degrees of freedom equal to the number of estimated loadings (p) for a particular item. The comparison test was analogous to the univariate test for difference in two means, but

extended to the multivariate case. First, a vector of differences between the estimated factor loadings of two subpopulations was computed. These differences followed a multivariate normal distribution with each element of the vector having an expected value of zero under the null hypothesis. Second, the variance-covariance matrix of the differences was formed by adding the variance-covariance matrices (similar to equation 5 but for only two subpopulations). Finally, the test statistic was calculated by pre- and post-multiplying the inverse of the variance-covariance of the differences ($\underset{\sim}{A}$) by the vector of differences ($\underset{\sim}{d}_i$):

$$(9) \quad \chi^2_i = \underset{\sim}{d}_i' \underset{\sim}{A}^{-1} \underset{\sim}{d}_i \cdot$$

with p degrees of freedom.

These comparisons identified those subpopulations across which the items had inconsistent factor loadings and accordingly the groups for which equivalence of measurement did not hold.

It is necessary to digress at this point to reflect on the number of statistical tests that were being made for a particular analysis and the accompanying possibility of committing type I errors. In testing the equality of subscale factor loadings across subpopulations, a test statistic was calculated for <u>each</u> subscale. In conducting so many tests, the likelihood of identifying a subscale as biased when in fact it was not, had increased. Therefore Wilks' lambda criterion served as a liberal statistic in that the possibility of overidentifying the true number of biased subscales definitely existed. At this point, the object was merely to identify the possible biased subscales, so committing overidentification errors was preferable to underidentification errors. The alpha level was not corrected for the number of tests. Thus Wilks' lambda served as a <u>preliminary</u> test to identify biased

subscales. As stated earlier, the SIFASP of the expanded biased subscale and the appropriate subscales results in a confirmation of Wilks' lambda. It was because of this confirmatory analysis that such liberal precautions were taken in the previous analysis. Having identified the specific items of the biased subscale for which the factor loadings were inconsistent across subpopulations, pairwise comparisons were made of the estimated loadings across the groups. For the case of three groups, it was necessary to adjust the alpha level ($\alpha/3$) to account for the increased probablity of committing type I errors.

RESULTS

Subscale Formation and Preliminary Analyses

Twenty-one subscales were formed from the sixty-one items of the ${}_{I}P_{I}$ on the criterion of item content. The correlations between the item responses and the corresponding propensity variables, $\rho(X_i, L_i)$, for the subscales showed a marked increase in the responses more nearly approaching normality, as opposed to the distributional characteristics of the individual items. While the item correlations ranged in value from .84 to .95, the creation of subscales improved the correlations to a range of .94 to .99. Thus the ${}_{I}P_{I}$ did yield item responses that were sufficiently normal and accordingly the items could be analyzed in their own right. But the analysis centered on subscales for reasons discussed earlier and the possibility for regressing back to the item level was definitely warranted.

The parallel analysis of the seven subscales confirmed the postulated seven factor structure of the ${}_{I}P_{I}$. Both real and generated psuedo-random eigenvalues resulting from principal axes factor analysis are presented in Table 1. The plot of the eigenvalues of real and generated pseudo-random variables intersected for eight common factors, implying that only seven

factors should be retained. The factor analysis for the seven factor solution provided the starting values for the Λ_y, Φ, and Θ^2 matrices which will be estimated by LISREL's maximum likelihood procedure.

Having fixed the diagonal elements of Φ and $m-1$ elements in each column of Λ_y to insure a unique solution, the maximum likelihood procedure confirmed the seven factor solution of the preliminary factor analysis. Table 1 presents the χ^2 value for each solution of one to eight factors. Although the χ^2 statistic is generally not a sensitive measure of the adequacy of the factor model, it did yield a nonsignificant result for the eighth factor ($p < .064$). The conclusion followed from this test that only seven factors were necessary to explain the interrelatedness of the subscales. The other measures of goodness of fit likewise verified the seven factor solution. The Tucker-Lewis coefficient showed a steady increase until the eighth factor, for which the increment was only .02. The other three indices, GFI, AGFI, and RMR, followed a similar pattern of lack of difference between the seventh and eighth factors.

This seven factor solution was further analyzed to determine the nonsignificant factor loadings. Since LISREL provides t-statistics for parameters that were estimated, a series of analyses was conducted which fixes at zero all estimates less than a certain t-value. Table 2 shows the results from the analyses in which the t-value was incremented by .5 for each factor solution. The model with the minimum number of fixed parameters was the one with m^2 or 49 parameters fixed. This model with 84 df and a χ^2 of 159.33 served as the comparison model against which all changes in df and χ^2 were determined. Using the ratio of change in χ^2 to change in degrees of freedom as an indication of fit between the factor model and the observed data, the analysis for an upper limit of t equal to 2.5 resulted in a ratio of 2.29. While this was a moderately large ratio, this analysis did not greatly differ on the other goodness of fit indices for the analysis of t equal 2.0. On the

other hand, fixing an additional four parameters at zero for t equal 3.0 resulted in an unacceptable ratio of 2.51.

To improve the fit of the model at the 2.5 t level, the modification indices were computed for the 122 fixed parameters. This index served as a means of identifying parameters that would significantly increase the fit of the model if relaxed (not constrained to equal zero). Distributed as a χ^2 with one df, the index identified three parameters with values greater than 11.00. Repeating the analysis but with these three parameters not fixed resulted in a ratio of 1.857, see the refinement section of Table 2. This process of determining the largest modification index, freeing the parameter, and calculating the χ^2 associated with the new model was repeated until the decrease in the χ^2 between models was no longer significant ($\chi^2_1 = 6.64$ for alpha = .01). The model with 116 parameter fixed satisfied this criteria in that freeing an additional parameter only resulted in a χ^2 change of 5.21. The ratio of changes in χ^2 to df associated with this model was 1.517, a marked improvement over the 2.29 value for the 122 fixed parameter model. Table 3 presents the estimated factor loadings and uniquenesses for the optimal factor solution. There was a clear and compelling simple structure for the total sample of 838 subjects. Table 4 presents the estimated factor variance-covariance matrix. Since the variances were fixed at unity, the matrix can be interpreted as the factor correlation matrix. These three matrices served as the foundation for all subsequent analyses.

SIFASP Analysis of Subscales

Because the units of measurement were arbitary for the subscales, the first step in the SIFASP was to rescale the variance-covariance matrices for each group according to equations 2-4. Box's test for equality of the three matrices resulted in a significant F ($p < .01$) implying that the matrices

differed. Since the null hypothesis of equal matrices was rejected, a SIFASP was conducted to determine which component of the variance-covariance matrices was the cause of the discrepancy.

Separate maximum likelihood factor solutions for the subscales were obtained for each subpopulation. Each analysis was based on the seven factor solution with the values in Tables 3 and 4 serving as the starting values. The parameters fixed at zero, one, or a constant were still fixed in these analyses. It was necessary to determine whether the common elements fixed at zero across the groups were actually zero in all subpopulations. The SIFASP analyses might not be able to detect factor loadings that were nonzero in some subpopulations while they were fixed at zero. To determine if the fixed parameters had any effect on the analysis, the modification indices for these parameters were examined for significance. Only three of the parameters resulted in significant modication indices ($.01 < p < .05$) for the three groups. All three parameters were from the student group, although all of the goodness of fit indices indicated that the factor model better fit the student data than it did the management data, see Table 5. Since there was no inordinate difference in the goodness of fit indices, the analyses continued as before with the same parameters fixed across all groups.

The χ^2 for the maximum likelihood factor solutions for each of the three groups is shown in Table 5. Since the groups were independent, the associated χ^2 values were independent and their sum (624.72) represented a χ^2 with 453 df. This χ^2 resulted from letting Λ_y, Φ, and Θ^2 vary across subpopulations, and is referred to as the unconstrained model.

Next, a simultaneous factor analysis was conducted in which both Λ_y and Θ^2 were constrained to be identical across the three groups. The Φ matrix was completely free to be estimated and thus an additional m nonzero elements in Λ_y were fixed. Fixed and free parameters (except for the above mentioned

changes) were set according to the values in Tables 3 and 4. The χ^2 for this constrained model was 816.02 for 557 df, an increase in the χ^2 of 191.30 for additional 104 df. This increase in χ^2 was significant ($p < .01$) but the ratio of change in χ^2 to change in df was relatively moderate, 1.84.

A second simultaneous factor anlaysis was conducted but with only Λ_y constrained across the groups. Both Φ and Ψ^2 were free to vary across the subpopulations. Imposing the restraint of equal factor loadings across the three groups resulted in an increase of 83.78 for χ^2 and 62 additional df. This result was marginally significant ($.025 < p < .05$) but the χ^2/df ratio was very small, implying that a different Λ_y matrix for each subpopulation better approximated the data than did some overall Λ_y matrix for the three groups. Table 6 presents the factor loading matrices for the three subpopulations. Both the GFI and RMR indices indicated an improvement in the quality of fit as Ψ^2 was freed to vary across subpopulations. Of course, the model in which only Λ_y was invariant showed better fit because this was more general (less parameters was constrained) than the model in which both Λ_y and Ψ^2 were constrained. Thus it was necessary to compare the models with respect to their number of constrained parameters as reflected in the change in df. A ratio of change in χ^2 to change in df was an indication of model fit across models. Although the increase in χ^2 was significant for constraining Λ_y, the ratio of changes, 1.38, was relatively small.

Finally, a simultaneous factor analysis was conducted in which only Ψ^2 was invariant across the subpopulations. The increase in χ^2 (107.52) was significant ($p < .01$) for its 42 df. Table 7 presents the estimated uniquenesses for both the constrained and unconstrained models. Again, GFI and RMR indicated this model as better fitting the data because there were less fixed parameters. Judged with respect to the change in df, the model yielded a

ratio that was large. Thus there was some evidence that the uniquenesses were not invariant across subpopulations. The results from this simultaneous analysis and the previous analysis suggested that the violations in equivalence of measurement across groups were more the result of differences among the unique variances than in differences in the factor pattern matrices. Although this was the case, the factor loadings were further analyzed to determine the subscales and subsequently the items of the ${}_{I}P_{I}$ which had inconsistent factor loadings across the groups.

In identifying the biased subscales, the estimated loadings were expressed in comparable metrics by using rescaled variance-covariance matrices and fixing the factor variances. The within and total SSCP matrices were calculated for each subscale and Wilks' lambda was computed as the quotient of their determinants. Distributed as an F, Wilks' lambda identified five biased subscales ($p < .01$): subscales 1, 6, 9, 13, and 19. Thus they were expanded to their original items for further analysis.

SIFASP Analysis of Items

The correlations between the item responses and the corresponding propensity variables, $\rho(X_i, L_i)$, ranged in value from .85 to .95 for the sixteen items that formed the biased subscales. Such values were sufficiently high to warrant the analysis being extended to the item level. In addition to these sixteen items, five subscales (one for each of the biased subscales) were included in the analysis as reference variables and to insure that at least three variables define each factor.

The procedure deviated at this point from the previous subscale analyses in that the SIFASP analysis was conducted first. Since it was already known that the factor loadings were not consistent across the three groups it was not necessary to conduct the simultaneous test. On the other hand, the SIFASP

analysis did provide the values at which to fix the factor variances in order to obtain a proper scaling of the factor loadings. The factor variances determined in the SIFASP analysis in which Λ_y was invariant but R^2 free, ranged in value from .818 to 1.218. Fixing the diagonal of each Φ matrix at these determined values as well as using the rescaled variance-covariance matrices for each group insured that the estimated factor loadings were comparable. The starting values for the Λ_y, R^2, and Φ matrices in the separate maximum likelihood solutions were provided by the estimates generated in the SIFASP analysis that identified the biased subscales. The subscale estimate was repeated for each item of the biased subscale.

The estimates of the factor loadings of the sixteen items and their standard errors are presented in Table 8 for each subpopulation. The standard errors were relatively consistent across the groups, generally ranging in value from .06 to .12. The standard errors appeared to be comparatively large for the first factor of the management subpopulation and the last few items on the sixth factor for students. Such discrepancies may influence the test of differences between the loadings across the groups. It should be noted that the largest and smallest factor variances were for these exact factors and groups, implying that these loadings may not be accurately estimated.

Wilks' lambda was calculated for each of the sixteen items and the significance levels are presented in Table 9. The statistic could not be calculated for the three items for which only one factor loading was estimated. Under such circumstances no between SSCP matrix could be computed and therefore the lambda statistic was indeterminant for these items. For those items which lambda could be calculated, all but one item was identified as biased ($p < .05$). There seemed to be no consistent pattern in the item response-item propensity correlations, both high and low values occured for the biased items.

The pairwise comparisons of the estimated loadings across subpopulations

confirmed the findings of the lambda statistic except for item number three, see Table 9. In this instance, lambda indicated that the item loadings were inconsistent across the groups ($p < .018$) while the pairwise test concluded the opposite (smallest significance level is $p < .086$). The pairwise comparisons of loadings classified ten items as biased for alpha < .01 and an additional three items were biased for alpha < .05. Only three out of the sixteen items forming the five biased subscales were not classified as biased by the pairwise comparisons. The biased items did not consistently differentiate the three subpopulations. Approximately the same number of biased items were significant for each of the three comparisons: nine for academic-management, nine for academic-student, and seven for management-student.

DISCUSSION

The discussion section is divided into two topic areas. First, the ultimate result of this study, the identification of biased items of the ${}_{I}P_{I}$, is discussed and consistencies among the items are noted. Referring back to Table 9 may be helpful throughout the following discussion. Second, to provide the total framework within which to interprete the results, the limitations of the study are addressed. The maximum likelihood factor analysis procedure is evaluated with respect to identifying biased items in the context of this particular study.

Interpretation of Biased Items

The fourteen biased items identified by the pairwise comparisons were examined for consistencies of content, format, and meaning. Each subscale had at least two biased items so patterns among the items will be identified along these dimensions.

The first set of biased items were on the task involvement scale of the $_{I}P_{I}$. The two items, "understood something for the first time" and "experienced adventure/novelty" reflect a theme of unique or new experiences. Both items differentiated the academic-student subpopulations. Such a result is indicative of the fact that students lack the vast array of experiences and the vigor in the search for understanding that motivates the academic profession. The academic sample of this study was largely composed of research oriented faculty members who are constantly seeking to understand and develop new ideas and theories. The "understood something for the first time" item likewise distinguished the academics from the management subpopulation. Again, academics would seem to dominate in this area while management would prefer to operate on a more routine and structured type of schedule. Note that these two groups did not differ on the "experience adventure/novelty" item, thus there was equivalent measurement across the two groups for this item. The third item of the biased subscale, "like a challenge" is not consistent with this novelty experience theme and accordingly is not identified as biased. Only one other item of the task involvement scale might be compatible with the described theme: "like to try something new." Whether the shift in tense affected the meaning of the item is unknown although the subscale for which this item is a member was not found to be biased.

Subscale six also had two biased items but were from the striving for excellence scale of the $_{I}P_{I}$. The two items pertain to the self improvement aspects of excellence: "improve my skills" and "compete against my self." The most significant distinction was between the academic and management groups for the "compete against myself" item. This item also differentiated the management and student groups. The idea of internal competition, setting high standards and working hard to achieve these goals is a constant issue confronting management. On the other hand, academic and students are not so

goal oriented. Their competition is often not with oneself but rather with the quest for knowledge. The "work hard to improve my skills" item only differentiated the academic and student subpopulations. This item reflects the basic mission of students. They are constantly in an environment that demands their persistence and dedication to work hard for the purpose of self improvement. Conversely, academics generally have expertise in a specific area and accordingly do not think in the same terms of working for self improvement as students do. The other items of the excellence scale are generally defined with respect to an individual's work or job and are not as personal in nature as the two biased items.

Both biased items forming the ninth subscale were from the competition scale of the $_{I}P_{I}$. The items reflected the drive to win and to be number one. "Like to win" distinguished students from both management and academics. The "need to be number one" item only differentiated the students from the academic. From these results it might be postulated that both management and academics are in highly competitive fields, at least more so than students. Whatever the interpretation, the conceptualization of the two items is inconsistent across the groups and accordingly the groups assign different meanings to the items as reflected in their responses.

All three biased items for the thirteenth subscale, from the affiliation scale, differentiated the management from the academics and students. Each of the items dealt with the concept of friendship: "in the company of friends," "loyalty to my friends," and "be friendly." It is difficult to provide any substantive reasoning for such a discrepancy but it can be concluded that the friendship theme is interpreted differently by management as compared to academics and students. Academics and student do not significantly differ in their views of these items. Two other items of the affiliation scale include the word "friends" or "friendly" but were not found to be biased. One of these

items is negatively worded ("I am suspicious of friendly people") and the other may be highly socially desirable or present a paradox to the respondent which may only elicit positive responses ("I am able to show warm affection to my friends").

The final set of biased items were from the recognition scale of the $_{I}P_{I}$. The four biased items of this scale, more so than any other set of items, differentiated all three groups. The items reflected a theme of the need for personal inspiration from others: "encouragement of others," "people tell me that I did well," "respect from coworkers," and "pleased someone important." The major difference between these four items and the other items forming the recognition scale is that all of the biased items mention a need for interaction with other people especially in terms of receiving recognition and respect. The other items of the scale, while also relevant to the need for recognition, do not allude to this interaction with colleagues or coworkers. Therefore the pairwise comparisons showed that each of the four items did not provide equivalent measurement across the subpopulations, implying that the groups had basically different conceptions associated with these items.

Assessment of the Maximum Likelihood Factor Analysis Procedure for the Detection of Item Bias

The maximum likelihood factor analysis procedure described and implemented in this study provided a particularly accomodating and statistically rigorous means of detecting item bias in attitudinal instruments. This section will evaluate the procedure as it was applied to the $_{I}P_{I}$ for the purpose of assessing the equivalency of measurement across several populations. The $_{I}P_{I}$ is representative of most attitudinal measures in that it is distinctly multidimensional, only moderate in length, and possesses a fixed number of response alternatives. In this regard, the generalizations resulting from this

study concerning the use of maximum likelihood factor analysis should extend to other attitudinal instruments. The following discussion is basically a step by step commentary of the maximum likelihood procedure, highlighting both its strengths and weaknessess.

The formation of the three subpopulations that were investigated was determined on the basis of similarity of profession. Although the groups may possibly have been heterogenous within themselves, if larger samples were available, comparisons could have been made on a more refined level. Despite the lack of such pure samples, the procedure was still able to function with subpopulations numbering only slightly more than 250 subjects. In fact, LOGIST (Wood, Wingersky, and Lord, 1976) did not converge after fifteen iterations even though the entire sample of 840 subjects was used. Whether this was the result of small sample size, the violation of the unidimensionality assumption, or some combination of the two, it further reinforced the applicability of the maximum likelihood technique for attitudinal measures.

An important limitation of the study involved the normality requirements for the factor model -- the subscales and especially the items were neither normally distributed nor continuous. Although the $\rho(X_i, L_i)$ values were relatively high for the subscales and acceptable for the items, this was not a definite indication that the variables were normally distributed. Such a violation of the assumption for maximum likelihood estimation may result in distorted estimates for the components of the factor model. Accordingly, if the factor model was poorly estimated the test statistics were inflated and were no longer an accurate indication of the goodness of fit by the factor model.

All analyses of the subscales, determining the optimal factor solution and identifying the biased subscales by the SIFASP analyses, quickly identified unique factor solutions, usually within four to six iterations. Such alacrity

in achieving the solution was likely the result of extremely good starting values from the preliminary factor analysis and high $\rho(X_i, L_i)$ values for all the subscales. While such quality in estimation of the solutions was evident, this did not address the concern of the quality of fit by the factor model. The residuals, normalized residual, and the modification indices of the subscales all indicated that the factor model more than adequately fit the data. In summary, the maximum likelihood procedure readily accomodated the analyses of subscales.

The quality of estimation and fit by the factor model did not totally extend to the item level. Estimation of factor model parameters for the items took as many as twenty-two iterations before convergence was reached for the management subpopulation. Such an outcome may result from a host of possible effects. First, the distribution of the item responses were less likely to approach normality. Second, the estimates for the biased subscales were merely repeated for all items as its starting value, maybe these values were not realistic for all items forming the subscales. Third, the same parameters that were fixed at zero for the subscales analyses were again fixed on the item level. It was possible that the parameters for the items were not consistent across the groups as the subscales parameters were. The modification indices for the three groups indicated that the fit of the model for the items would significantly increase if these parameters were set free.

Finally, the large number of statistical tests being made to determine biased subscales and biased items must be taken in account. Such testing may possibly lead to a number of significant results occurring by chance. In the initial analyses, steps were taken to minimize the number of "true" biased subscales being misclassified by overidentifying the number of subscales considered biased. Such precaution may result in an inordinate number of items being classified as biased due to this overidentification of subscales.

In conclusion, the results in this study showed that consistencies within the biased items can be identified. The determined themes of the items indicated that the biased items were quite homogeneous in their content. Generally the biased item theme was <u>not</u> representative of other items on the $_{I}P_{I}$ scale. Therefore it can be concluded that thirteen items of the $_{I}P_{I}$ did not provide equivalent measurement across the three subpopulations of this study with respect to the identified biased items. The maximum likelihood factor analysis procedure provided many advantages over the classical test theory and item response theory approaches for the detection of item bias in attitudinal instruments. Despite its limitations, the technique was readily able to accomodate multidimensional data sets that were polychotomously scored. Also the procedure provided a very rigorous statistical test that is deeply rooted in factor analysis research and is a viable alternative to the more restrictive item response theory techniques.

(1) Test the equality of variance-covariance matrices across subpopulations.

$$H_o: \hat{\Sigma}_1 = \hat{\Sigma}_2 = \hat{\Sigma}_3$$

Test statistic = Box's F.

	Significant	Nonsignificant
Conclusion	Matrices differ.	Equivalence of measurement.
Action	Test invariance properties of Λ_y & Ψ^2.	End analysis.

(2) Test effect of constraining Λ_y and Ψ^2 to being equal across subpopulations.

$$H_o: \Lambda_{y1} = \Lambda_{y2} = \Lambda_{y3} \text{ and } \Psi_1^2 = \Psi_2^2 = \Psi_3^2$$

$$\text{Test statistic} = \frac{\chi^2_{\Lambda_y\Psi^2} - \chi^2_{\text{unconstrained}}}{df_{\Lambda_y\Psi^2} - df_{\text{unconstrained}}}.$$

	Significant	Nonsignificant
Conclusion	Λ_y or Ψ^2 is not invariant.	Equivalence of measurement.
Action	Free Ψ^2 and retest.	End analysis.

(3) Test effect of constraining Λ_y to be equal across subpopulations.

$$H_o: \Lambda_{y1} = \Lambda_{y2} = \Lambda_{y3}$$

$$\text{Test statistics} = \frac{\chi^2_{\Lambda_y} - \chi^2_{\text{unconstrained}}}{df_{\Lambda_y} - df_{\text{unconstrained}}}.$$

	Significant	Nonsignificant
Conclusion	Λ_y is not invariant, equivalence of measurement does not hold.	Equivalence of measurement but Ψ^2 is not invariant.
Action	Identify subscales across which Λ_y is invariant. Repeat analysis but on item level.	End analysis.

Figure 1. Progression of SIFASP analysis.

Table 1

Maximum Likelihood Factor Analysis of the Corelation Matrix of Twenty-One Subscales for the $_{I}P_{I}$

Number of Factors	df	χ^2 (prob)	Tucker-Lewis	Goodness of Fit Index	Adjusted Goodness of Fit Index	Root Mean Square Residual	Eigenvalues * Real	Random
1	189	3605.47 (.000)	.4120	.650	.572	.123	4.963	.317
2	169	2488.46 (.000)	.5532	.740	.645	.092	1.952	.264
3	150	1669.63 (.000)	.6699	.813	.712	.073	1.266	.222
4	132	1115.14 (.000)	.7571	.869	.770	.049	1.188	.190
5	115	726.25 (.000)	.8265	.908	.815	.038	.647	.163
6	99	402.06 (.000)	.9000	.947	.877	.029	.515	.139
7	84	159.33 (.000)	.9707	.980	.945	.014	.377	.117
8	70	88.78 (.064)	.9912	.989	.965	.009	.095	.095

* The real eigenvalues are determined by a principal axes factor analysis and the random eigenvalues are determined by a parallel analysis.

Table 2

Determination of Optimal Factor Solution for Twenty-One Subscales

Number of Parameters Fixed	Upper t-value Limit	df	X^2	GFI	AGFI	RMR	Change in df	Change in X^2	Ratio
49	-	84	159.33	.980	.945	.014			
81	1.0	116	170.48	.977	.955	.016	32	11.15	.348
98	1.5	133	219.65	.965	.939	.023	49	60.32	1.23
116	2.0	151	298.84	.947	.919	.030	67	139.51	2.08
122	2.5	157	326.35	.940	.911	.032	73	167.02	2.29
126	3.0	161	352.60	.929	.902	.039	77	193.27	2.51
Refinement Stage *									
119	-	154	289.31	.951	.926	.028	70	129.98	1.86
118	-	153	279.73	.953	.929	.026	69	120.40	1.75
117	-	152	268.67	.955	.932	.025	68	109.34	1.61
116	-	151	260.99	.958	.936	.025	67	101.66	1.52
115	-	150	255.78	.961	.939	.024	66	96.45	1.46

* Refinements are made according to the largest modification index for the model yielding changes greater than 6.64 (alpha = .01, 1 df).

Table 3

Estimated Factor Loadings and Uniqueness for the $_{I}P_{I}$ Subscales
Optimal Factor Solution Determined by Maximum Likelihood Factor Analysis

	Factor							
Subscale	I	II	III	IV	V	VI	VII	$\underset{\sim}{D}^2$
1		.69 (04)			-.16 (04)	.24 (04)		.44 (03)
2		.55 (04)				.19 (03)	.12 (04)	.50 (03)
3		.86 (03)						.25 (04)
4							.75 (04)	.44 (04)
5						.12 (04)	.51 (04)	.70 (04)
6							.70 (04)	.52 (04)
7					.85 (03)			.29 (03)
8					.53 (03)			.72 (04)
9	-.17 (05)		.13 (04)		.87 (06)		-.14 (04)	.36 (04)
10			.76 (04)				.04 (04)	.40 (03)
11			.80 (03)					.36 (03)
12			.80 (03)	-.13 (03)				.40 (03)
13	-.12 (04)			.70 (04)			-.14 (03)	.54 (04)
14				.63 (04)				.61 (04)
15				.69 (04)				.52 (04)
16	.67 (04)				.19 (04)			.37 (03)
17	.80 (03)							.37 (03)
18	.79 (03)							.37 (03)
19		-.23 (04)	-.13 (04)	.47 (04)		.63 (04)		.53 (04)
20						.77 (03)	.09 (04)	.36 (03)
21						.85 (03)		.29 (03)

Note: N = 838; standard errors of parameter estimates are within parentheses, decimals have been left off; nonsignificant factor loadings are left blank and are fixed at zero for all further analyses.

Table 4

Estimated Factor Variance-Covariance Matrix for the Twenty-One Subscales

	Factor						
Factor	I	II	III	IV	V	VI	VII
I	1.00						
II	.04 (04)	1.00					
III	.31 (04)	.20 (04)	1.00				
IV	-.15 (05)	.25 (04)	.25 (05)	1.00			
V	.53 (04)	.33 (04)	.35 (04)	-.04 (05)	1.00		
VI	.58 (03)	.37 (04)	.42 (04)	.06 (05)	.59 (03)	1.00	
VII	.13 (04)	.48 (04)	.48 (04)	.21 (05)	.37 (04)	.28 (05)	1.00

Note: Standard errors of parameter estimates are within parentheses, decimal points are omitted.

Table 5

Maximum Likelihood Factor Analyses and Simultaneous Factor Analyses for Three Subpopulations

MAXIMUM LIKELIHOOD FACTOR ANALYSIS (Unconstrained Model)

Group	N	χ^2_{151}	GFI	AGFI	RMR
Academic	322	225.18	.903	.866	.035
Management	248	205.65	.900	.847	.040
Students	268	193.89	.902	.850	.040
		624.72			

SIMULTANEOUS FACTOR ANALYSIS (Constrained Models)

Goodness of Fit Indices	Invariant Matrix		
	Λ_y & Φ^2	Λ_y	Φ^2
χ^2, df	816.02, 557	708.50, 515	732.24, 495
$\Delta\chi^2$, Δdf	191.30, 104	83.78, 62	107.52, 42
Ratio	1.839	1.351	2.560
GFI			
Academic	.878	.884	.906
Management	.865	.876	.888
Students	.856	.871	.885
RMR			
Academic	.043	.041	.036
Management	.048	.046	.040
Students	.054	.052	.047

Table 6

Estimated Factor Loadings of Subscales Resulting From Separate Maximum Likelihood Factor Analyses of Rescaled Variance-Covariance Matrices For Three Subpopulations

Group	Subscale	Factor I	II	III	IV	V	VI	VII
Academic	1		.71 (06)			-.17 (04)	.16 (07)	
	2		.49 (07)				.18 (05)	.20 (06)
	3		.90 (05)					
	4							.63 (06)
	5						.12 (06)	.42 (06)
	6							.82 (06)
	7					.88 (05)		
	8					.61 (05)		
	9	-.12 (07)		.12 (06)		.84 (07)		-.07 (05)
	10			.79 (06)				.03 (06)
	11			.80 (05)				
	12			.83 (05)	-.11 (050			
	13	.18 (06)			.71 (06)			-.08 (06)
	14				.70 (06)			
	15				.72 (06)			
	16	.74 (06)				.14 (06)		
	17	.81 (05)						
	18	.75 (05)						
	19		-.21 (06)	-.16 (07)	.38 (06)		.68 (08)	
	20						.74 (05)	.12 (05)
	21						.82 (05)	
Management	1		.75 (07)			-.16 (08)	.22 (07)	
	2		.56 (07)				.15 (06)	.15 (07)
	3		.90 (06)					
	4							.78 (07)
	5						.16 (07)	.47 (08)
	6							.62 (07)
	7					.82 (06)		
	8					.47 (07)		
	9	-.10 (10)		.23 (09)		.86 (14)		-.35 (13)
	10			.80 (08)				.01 (08)
	11			.80 (06)				
	12			.71 (06)	-.03 (06)			
	13	.05 (07)			.61 (08)			-.09 (08)
	14				.53 (07)			
	15				.68 (07)			
	16	.61 (07)				.1 (07)		
	17	.84 (06)						
	18	.81 (06)						
	19		-.26 (08)	-.23 (09)	.54 (08)		.66 (10)	
	20						.70 (07)	.11 (07)
	21						.84 (06)	
Students	1		.61 (07)			-.13 (07)	.34 (08)	
	2		.57 (07)				.21 (06)	.07 (07)
	3		.80 (06)					
	4							.78 (06)
	5						.13 (06)	.54 (07)
	6							.70 (06)
	7					.83 (06)		
	8					.44 (06)		
	9	-.33 (12)		.12 (07)		.98 (13)		-.13 (09)
	10			.68 (07)				.13 (06)
	11			.79 (06)				
	12			.84 (07)	-.24 (06)			
	13	.09 (06)			.78 (07)			-.19 (07)
	14				.63 (06)			
	15				.68 (06)			
	16	.67 (07)				.21 (07)		
	17	.72 (06)						
	18	.80 (06)						
	19		-.22 (07)	-.06 (06)	.51 (07)		.63 (07)	
	20						.83 (05)	.06 (05)
	21						.88 (05)	

Note: Standard errors of parameter estimates are within parentheses, decimal points are omitted.

Table 7

Estimated Uniquenesses of Subscales From Constrained And Unconstrained Simultaneous Factor Analyses of Three Subpopulations

Subscale	Constrained $\underset{\sim}{D}^2$	Unconstrained $\underset{\sim}{D}^2$		
		Academic	Manage	Student
1	.45 (03)	.44 (04)	.43 (05)	.43 (05)
2	.51 (03)	.47 (04)	.49 (05)	.55 (06)
3	.23 (04)	.17 (04)	.15 (05)	.47 (07)
4	.47 (04)	.51 (06)	.34 (05)	.56 (07)
5	.70 (04)	.64 (06)	.74 (07)	.74 (07)
6	.50 (04)	.41 (05)	.58 (07)	.50 (07)
7	.28 (03)	.23 (04)	.28 (04)	.35 (06)
8	.73 (04)	.65 (06)	.66 (06)	.88 (08)
9	.40 (04)	.35 (05)	.44 (06)	.40 (06)
10	.41 (03)	.35 (04)	.38 (05)	.50 (06)
11	.36 (03)	.30 (04)	.31 (04)	.51 (06)
12	.41 (03)	.31 (04)	.51 (06)	.43 (06)
13	.52 (04)	.51 (06)	.57 (07)	.49 (06)
14	.60 (04)	.54 (05)	.65 (07)	.62 (07)
15	.54 (04)	.52 (06)	.51 (06)	.59 (07)
16	.38 (03)	.32 (04)	.46 (05)	.38 (04)
17	.37 (03)	.34 (04)	.36 (05)	.41 (05)
18	.38 (03)	.38 (04)	.39 (05)	.37 (06)
19	.55 (04)	.48 (05)	.66 (07)	.52 (06)
20	.37 (03)	.37 (04)	.36 (04)	.37 (05)
21	.28 (03)	.27 (04)	.26 (04)	.31 (05)

Note: Standard errors of parameter estimates are within parentheses, decimal points are omitted.

Table 8

Estimated Factor Loadings of Items Resulting From Separate Maximum Likelihood Factor Analyses of Rescaled Variance-Covariance Matrices For Three Subpopulations

Group	Item	Factor						
		I	II	III	IV	V	VI	VII
	1		.29 (07)			-.25 (08)	.35 (06)	
	2		.27 (06)			-.28 (08)	.36 (06)	
	3		.70 (07)			.22 (07)	-.11 (06)	
	4							.46 (06)
	5							.46 (06)
	6							.34 (06)
	7	-.34 (12)		.07 (06)		.89 (11)		-.14 (08)
Academic	8	-.47 (14)		.04 (07)		1.01 (13)		-.02 (08)
	9	.27 (08)			.48 (07)			-.29 (07)
	10	-.01 (08)			.54 (06)			-.08 (07)
	11	-.21 (07)			.35 (06)			.03 (06)
	12	-.07			.36 (06)			.05 (07)
	13		-.39 (08)	-.08 (06)	.60 (06)		.30 (05)	
	14		-.26 (07)	-.01 (05)	.25 (06)		.50 (05)	
	15		-.29 (07)	-.04 (06)	.34 (07)		.42 (05)	
	16		-.22 (07)	.02 (06)	.43 (08)		.30 (05)	
	1		.35 (08)			-.30 (09)	.27 (10)	
	2		.54 (08)			-.25 (09)	.13 (10)	
	3		.73 (07)			.06 (08)	.01 (08)	
	4							.53 (08)
	5							.72 (08)
	6							.37 (08)
	7	-.17 (30)		.10 (09)		.75 (09)		-.18 (09)
Management	8	-.84 (32)		.07 (09)		.95 (10)		.07 (09)
	9	.98 (27)			.53 (10)			-.17 (11)
	10	.67 (26)			.62 (10)			-.03 (10)
	11	.72 (26)			.50 (10)			-.14 (10)
	12	.40 (25)			.44 (10)			-.06 (10)
	13		-.50 (10)	-.18 (10)	.55 (09)		.63 (10)	
	14		-.54 (09)	-.02 (09)	.20 (09)		.75 (10)	
	15		-.63 (10)	-.30 (10)	.42 (09)		.87 (11)	
	16		-.49 (10)	-.26 (10)	.43 (09)		.72 (10)	
	1		.50 (09)			-.45 (11)	.39 (10)	
	2		.52 (09)			-.40 (11)	.31 (10)	
	3		.73 (08)			.29 (09)	-.13 (08)	
	4							.63 (07)
	5							.52 (07)
	6							.27 (07)
	7	-.22 (18)		.27 (11)		1.04 (16)		-.55 (14)
Student	8	-.46 (21)		.26 (12)		1.26 (19)		-.53 (16)
	9	.33 (11)			.41 (09)			-.23 (09)
	10	-.01 (11)			.44 (08)			.03 (08)
	11	-.14 (11)			.48 (08)			.08 (08)
	12	-.10 (11)			.28 (08)			-.01 (08)
	13		-.88 (15)	-.54 (14)	.83 (12)		1.01 (17)	
	14		-.71 (13)	-.35 (12)	.24 (11)		1.08 (14)	
	15		-.75 (15)	-.53 (14)	.81 (12)		1.02 (17)	
	16		-.44 (13)	-.32 (12)	.70 (11)		.69 (14)	

Note: Standard errors of parameter estimates are within parentheses, decimal points are omitted.

Table 9

Significance Levels for Wilks' Lambda and Pairwise Comparison of Item Factor Loadings Across Three Subpopulations

Subscale	Item	$\rho(X_i, L_i)$	Wilks' Λ	Acad-Man	Acad-Stud	Man-Stud
	1	.852	.018	.148	.289	.086
1	2	.890	.019	.010 **	.039 *	.183
	3	.924	.026	.175	.015 *	.095
	4	.903	#	.151	.017 *	.105
6	5	.940	#	.002 **	.172	.018 *
	6	.941	#	.255	.148	.114
9	7	.935	.002	.221	.015 *	.003 **
	8	.951	.003	.285	.026 *	.077
	9	.853	.005	.010 **	.193	.034 *
13	10	.924	.143	.117	.247	.089
	11	.930	.017	.002 **	.157	.004 **
	12	.931	.029	.006 **	.249	.036 *
	13	.847	.010	.002 **	.018 *	.074
19	14	.906	.003	.010 **	.001 **	.048 *
	15	.937	.002	.013 *	.000 **	.079
	16	.941	.004	.001 **	.001 **	.037 *

Wilks' lambda can not be calculated for this subscale because only one factor loading is estimated and accordingly no between SSCP matrix can be computed.

* Significant at alpha < .05.
** Significant at alpha < .01.

REFERENCES

Braskamp, L. A. and Maehr, M. L. Development of the Inventory of Personal Investment. Paper presented at the annual meeting of the American Educational Research Association, Montreal, April, 1983.

Box, G. E. A general distribution theory for a class of likelihood criteria. Biometrika, 1949, 36, 317-346.

Drasgow, F. and Dorans, N. J. Robustness of estimators of the squared multiple correlation and squared cross-validity coefficient to violations of multivariate normality. Applied Psychological Measurement, 1982, 6, 185-200.

Drasgow, F. and Kanfer, R. Equivalence of psychological measurement in hetergeneous populations. Manuscript in preparation, 1983.

Humphreys, L. G. and Montanelli, R. G. An investigation of the parallel analysis criterion for determining the number of common factors. Multivariate Behavioral Research, 1975, 10, 193-205.

Ironson, G. H. and Subkoviak, M. J. A comparison of several methods of assessing item bias. Journal of Educational Measurement, 1979, 16, 209-225.

Joreskog, K. G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Joreskog, K. G. Some contributions to maximum likelihood factor analysis. Psychometrika, 1967, 32, 443-482.

Joreskog, K. G. and Sorbom, D. Advances in Factor Analysis and Structural Equation Models. Cambridge, MA: Abt Books, 1979.

Joreskog, K. G. and Sorbom, D. LISREL V: Analysis of Linear Structural Relationsips by the Method of Maximum Likelihood User's Guide. Chicago: National Educational Resources, 1981.

Kroonenberg, P. M. and Lewis, C. Methodological issues in the search for a factor model: Exploration through confirmation. Journal of Educational Statistics, 1982, 7, 69-89.

Lawley, D. N. A note on Karl Pearson's selection formulae. Proceedings of the Royal Society of Edinburgh, Section A, 1943-44, 62, 28-30.

Lord, F. M. Applications of Item Response Theory to Practicial Testing Problems. Hillsdale, NJ: Erlbaum, 1980.

Lord, F. M. A study of item bias using item characteristic curve theory. In Y. H. Poortingal (Ed.) Basic Problems in Cross-Cultural Psychology. Amsterdam: Swets and Zeitlinger, 1977.

McGaw, B. and Joreskog, K. G. Factorial invariance of ability measures in groups differing in intelligence and socioeconomic status. British Journal of Mathematical and Statistical Psychology, 1971, 24, 154-168.

Meredith, W. Rotation to achieve factorial invariance. Psychometrika, 1964, 29, 187-206.

Mulaik, S. A. The Foundations of Factor Analysis. New York: McGraw-Hill Book Company, 1972.

Scheumeman, J. A method of assessing bias in test items. Journal of Educational Measurement, 1979, 16, 143-152.

Tatsuoka, M. M. Multivariate Analysis: Techniques for Educational and Psychological Research. New York: John Wiley and Sons, 1971.

Thurstone, L. L. Multiple Factor Analysis. Chicago: University of Chicago Press, 1947.

Tucker, L. R. and Lewis, C. A reliability coefficient for maximum likelihood factor analysis. Psychometrika, 1973, 38, 1-10.

Wilks, S. S. Certain generalizations in the analysis of variance. Biometrika, 1932, 24, 471-494.

Wood, R. L., Wingersky, M. S., and Lord F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. Research Memorandum 76-6. Princeton, NJ: Educational Testing Service, 1976.

Wright, B. D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 91-116.